PheProb: probabilistic phenotyping using diagnosis codes to improve power for genetic association studies

Jennifer A Sinnott, Fiona Cai, Sheng Yu, Boris P. Hejblum, Chuan Hong, Isaac Kohane, Katherine P. Liao

PheProb: Probabilistic Phenotyping Using Diagnosis Codes to Improve Power for Genetic Association Studies

Jennifer A. Sinnott[1], Fiona Cai[2], Sheng Yu[3,4], Boris P. Hejblum[5], Chuan Hong[6], Isaac S. Kohane[7], Katherine P. Liao[8]

[1]*Department of Statistics, The Ohio State University, Columbus, OH, USA;*

[2]*Stuyvesant High School, New York City, NY, USA;*

[3]*Center for Statistical Science, Tsinghua University, Beijing, China;*

[4]*Department of Industrial Engineering, Tsinghua University, Beijing, China;*

[5]*Univ. Bordeaux, ISPED, Inserm BPH 1219, Inria SISTM, Bordeaux, France;*

[6]*Department of Biostatistics, Harvard University, Boston, MA, USA;*

[7]*Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA;*

[8]*Department of Medicine, Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, MA, USA*

Correspondence to:
Jennifer A. Sinnott
Cockins Hall
1958 Neil Ave.
Columbus OH, 43210, USA
Email: jsinnott@stat.osu.edu
Tel: +1-614-292-8110

Keywords: electronic health records, genetic association test, low-density lipoprotein, rheumatoid arthritis, mixture model, phenome-wide association study

Word count: 3974

ABSTRACT

Objective: Standard approaches for large scale phenotypic screens using electronic health record (EHR) data apply thresholds, such as ≥2 diagnosis codes, to define subjects as having a phenotype. However, the variation in the accuracy of diagnosis codes can impair the power of such screens. Our objective was to develop and evaluate an approach which converts diagnosis codes into a probability of a phenotype (PheProb). We hypothesized that this alternate approach for defining phenotypes would improve power for genetic association studies.

Methods: The PheProb approach employs unsupervised clustering to separate patients into two groups based on diagnosis codes. Subjects are assigned a probability of having the phenotype based on the number of diagnosis codes. This approach was developed using simulated EHR data and tested in a real world EHR cohort. In the latter, we tested the association between low density lipoprotein cholesterol (LDL-C) genetic risk alleles known for association with hyperlipidemia and hyperlipidemia codes (ICD-9 272.x). PheProb and thresholding approaches were compared.

Results: Among n=1,462 subjects in the real world EHR cohort, the threshold-based p-values for association between the genetic risk score (GRS) and hyperlipidemia were 0.126 (≥1 code), 0.123 (≥2 codes), and 0.142 (≥3 codes). The PheProb approach produced the expected significant association between the GRS and hyperlipidemia: p=0.001.

Conclusion: PheProb improves statistical power for association studies relative to standard thresholding approaches by leveraging information about the phenotype

in the billing code counts. The PheProb approach has direct applications where efficient approaches are required such as in Phenome-Wide Association Studies.

INTRODUCTION

Electronic health records (EHRs) contain a wealth of comprehensive patient information. When linked with genomic data, the combined information provides a powerful platform to study associations between genetic variants and a variety of diseases, disorders, and other conditions. Ideally, disease cases and controls in large, diverse populations would be identified automatically using data in the EHR, and linked to genetic markers assessed on collected blood samples [1]. However, there is currently a mismatch wherein the ability to extract accurate information about patient phenotypes from EHRs lags behind genotyping [2]. For example, although EHRs often contain diagnosis codes for specific diseases, the presence or absence of these codes is not perfectly correlated with presence or absence of the disease. Previous studies have successfully replicated established genetic associations by building phenotyping algorithms using both structured data such as diagnosis codes and unstructured data such as physicians' notes accessed by natural language processing [3–6]. A major limitation of these approaches is the requirement for labor intensive chart review to establish gold standard labels on a subset of cases. Such approaches are difficult to scale when multiple phenotypes are of interest.

In particular, with the increasing availability of large cohorts with linked EHR and genetic data, there is growing interest in screening for associations between a genetic marker of interest and a wide range of clinical phenotypes — that is, in performing a Phenome-Wide Association Study (PheWAS) [3,7–20]. In a PheWAS, developing highly accurate algorithms incorporating structured and unstructured

EHR data for each phenotype is infeasible. Instead, researchers typically rely on available structured data such as demographic information and International Classification of Diseases (ICD) billing and diagnosis codes (frequently ICD-9 or ICD-10, signifying the Ninth or Tenth Revisions of the Classification system, respectively). In the most common PheWAS approach, these codes are collapsed across time and stored as counts for each code for each individual. Tens of thousands of ICD codes can then be converted into a smaller number of phenotypes, as proposed in Denny et al. — the phenotype billing code counts created by this conversion are typically called *PheWAS codes* [9,15]. For each phenotype, "cases" are individuals with at least one relevant billing code and "controls" are individuals with zero codes (and sometimes without codes for other related phenotypes). With this approach, Denny et al. successfully replicated 4 of 7 known disease-SNP associations at the 0.05 level [9].

For the three associations that did not replicate in Denny et al., they noted that identifying cases as individuals with at least one relevant ICD-9 code was not a stringent enough definition; many of these "cases" did not have the disease when their medical records were manually reviewed. For these diseases, the "case" definition that was used had low positive predictive value (PPV). Indeed, the reliability of using presence of any relevant billing code as a proxy for presence of the phenotype varies by phenotype and health care system [21,22]. Subsequent studies proposed adaptations of this approach to improve the PPV, such as requiring at least two billing codes on two different days [13,20,23]; or a number of billing codes that differs depending on disease frequency [10,19]; or two, three, or four

billing codes in total [12,16–18,24]. While these more stringent definitions improve the PPV of the case identification, they can also reduce power by eliminating some true cases, which can be problematic for uncommon phenotypes. These studies identify a gap in knowledge regarding an efficient approach for determining the optimal threshold. Moreover, collapsing a billing code count into a binary case-control status, and not accounting for total healthcare usage that can vary dramatically among patients, may eliminate information that could better distinguish cases and controls.

In this study, we propose an automated approach for using diagnosis codes that avoids setting an arbitrary threshold on the number of codes required to define a "case" by instead converting the diagnosis codes into a probability of a phenotype through unsupervised clustering (PheProb). We will compare the performance of genetic association testing with PheProb-defined phenotypes and threshold-based phenotypes in simulated and real world EHR data. Using known associations, we hypothesize that the PheProb approach will demonstrate stronger genotype-phenotype associations than with the thresholding approach.

METHODS

Development of methods for PheProb

We used simulated data that mimicked the patterns of real EHR data, but with gold standard phenotype labels available, to compare the performance of PheProb with the performance of threshold-based methods. The dataset contained 2000 patients

with EHR data linked with genotype data; each patient had a single nucleotide polymorphism (SNP) value — 0, 1, or 2 minor alleles — and a single normally distributed clinical covariate. Disease status was generated so that the probability of having the disease depended on the value of the SNP and the covariate; we considered both weak (odds ratio [OR] =1.1), and moderate (OR=1.35) disease-SNP associations. We also varied disease prevalence: 20%, 10%, and 5%. The total number of billing codes for each patient was randomly generated, and the number of disease PheWAS codes was generated from a binomial distribution with sample size equal to the total number of billing codes and success probability dependent on the underlying disease status.

In the standard PheWAS, disease cases are assigned by thresholding the relevant billing codes using a cut-off, as was proposed in Denny et al. (2010) [9,15]. Thus, as a comparison for our approach we applied the standard PheWAS thresholds of $\geq 1, \geq 2$, and ≥ 3 ICD codes to define a disease case; individuals with no ICD codes were controls. To perform the genetic association test, we tested for association between the SNP and case-control status, with cases defined by the three thresholds; the three models are denoted by S_1, S_2, and S_3.

The proposed PheProb method applies a different approach to define disease status and perform the genetic association test. A diagram of the workflow of this two-step approach is provided in Figure 1.

In Step 1, we fit a mixture model to the disease-relevant billing code count variable, S, assuming two latent classes — cases ($Y = 1$) and controls ($Y = 0$). Specifically, we assume that there are two underlying classes, and that within each

class, S follows a binomial distribution, $P(S = s|C = c) = \binom{c}{s} {p_d}^s (1 - p_d)^{c-s}$, with parameters $n = c$, the total number of billing code counts, and $p_d = p_1$ if the patient is a disease case or p_0 if the patient is a control. Thus, the model accounts for the total healthcare utilization, as quantified by the total number of billing codes C, by interpreting the count of relevant billing codes S as a subset of C in these class-specific binomial models. The probability of belonging to the case population is denoted by φ; if φ were a single value, it would represent the underlying prevalence of the disease, but our proposed method benefits from allowing the prevalence to vary somewhat according to total health care utilization C. Thus, we let $\varphi(c, \alpha_0, \alpha_1) = g(\alpha_0 + \alpha_1 c)$, where the parameters $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$ are unknown and g is the logistic function. All parameters $(p_1, p_0, \alpha_0, \alpha_1)$ are estimated using the expectation-maximization (EM) algorithm, so that this model can be fit in an unsupervised manner and thus does not require any "gold standard" labelling of cases and controls [25–28]. After the model is fit and α_0, α_1, p_0, and p_1 are estimated, we can calculate the predicted probability that each patient is a disease case by Bayes rule:

$$\hat{\pi}_Y = \hat{P}(Y = 1|S = s, C = c) = \frac{\hat{\varphi}\binom{c}{s}{\hat{p}_1}^s (1-\hat{p}_1)^{c-s}}{\hat{\varphi}\binom{c}{s}{\hat{p}_1}^s (1-\hat{p}_1)^{c-s} + (1-\hat{\varphi})\binom{c}{s}{\hat{p}_0}^s (1-\hat{p}_0)^{c-s}}.$$

Here, we write $\hat{\varphi}$ as shorthand for $\varphi(c, \hat{\alpha}_0, \hat{\alpha}_1)$.

In Step 2, we test whether the SNP is associated with the probability of having the phenotype, and calculate a p-value quantifying the strength of that association. Essentially, we calculate the model's predicted probability, $\hat{\pi}_{Yi}$, for each individual as defined above, and we use $\hat{\pi}_{Yi}$ in place of a typical $0/1$ outcome in a

logistic regression model with the genetic marker G and any clinical variables such as age and gender included as covariates in the model. We fit the model using logistic regression estimating equations and calculate a robust variance estimate to use for testing whether there is any association between the disease and the genetic marker, controlling for clinical variables [29].

This approach allows the model connecting diagnosis codes and probabilities to differ disease-to-disease based on the features of the disease distributions. A more detailed description of the simulation settings and the statistical models and methods is available in the Web Appendix. Implementation is available in the PheProb R package, available from the authors on request.

Comparison of Existing Methods and PheProb

Simulated Data

For each simulated dataset, we applied the standard threshold-based genotype-phenotype association tests, S_1, S_2, and S_3. We also applied our proposed PheProb method. For comparision, we fit a model using the true disease status; while not feasible in practice, this was informative for benchmarking. Simulations were run 500 times for each setting in R [30]; the flexmix and gee packages were used in our simulations [25–27,29].

Example 1: Real-World Clinical Study and Hyperlipidemia

A group of control subjects from a previous EHR genetic study of lipids was used to test PheProb on real world data [31]. Briefly, the study consisted of 1,462 subjects

with clinical EHR data including demographics, diagnosis codes, and genetic data. In the prior study [31], we confirmed that patients carrying more LDL-C risk alleles, aggregated into a composite LDL genetic risk score (GRS), had higher LDL-C levels measured as part of routine care. LDL-C is the target of statins and is considered part of the causal pathway for cardiovascular disease. High LDL-C, or hyperlipidemia, corresponds to ICD-9 codes starting with 272. For this study, all such codes were collapsed into a count of the number of hyperlipidemia-related billing codes. We tested for association between the LDL GRS and the phenotype hyperlipidemia and compared the p-value for association when the phenotype was defined using standard threshold-based methods compared to the proposed PheProb method. All models were adjusted for age and sex.

Example 2: Partners Biobank and Rheumatoid Arthritis

We additionally studied a population where gold standard labels were available for the phenotype rheumatoid arthritis (RA). The Partners Healthcare Biobank comprises 14,985 subjects enrolled from 2011-2016 with both clinical EHR data and genetic data [32]. For a subset of 546 of these patients, chart review was performed to confirm the presence of the most common autoimmune inflammatory joint disease, rheumatoid arthritis (RA). We extracted data on two of the strongest genetic risk alleles for RA, rs9268839 in *HLA DRB1* and rs2476601 in *PTPN22*. The PheWAS codes 714 and 714.1 correspond to RA and other inflammatory polyarthropathies. Based on patients' counts of these codes, we tested whether *HLA*

DRB1 and *PTPN22* were associated with RA, adjusting for age, gender and race, using both the standard thresholding method and the PheProb approach.

For the subset of 546 individuals with true RA status known (from manual chart review), we sought to better understand how well the standard thresholding methods and the PheProb clustering step were separating cases and controls. To do this we estimated the false positive rate (FPR), the recall, the precision / PPV, the negative predictive value (NPV), and the F1 score for the threshold-defined case-control status, focusing on S_2, the most popular approach in the literature [12,16,18]. For PheProb, we estimated the same quantities, after thresholding the method's estimated probability at its mean to define a binary outcome.

RESULTS

Simulated Data Results

In Figure 2, the power to detect the weak (OR=1.1) and moderate (OR=1.35) genotype-phenotype associations was compared across all methods. In all settings, the PheProb approach outperformed the current standard threshold-based methods (S_1, S_2, S_3). For example, when the disease prevalence is medium and the association is moderate (OR=1.35, prevalence = 10%), the power to detect the association with PheProb is 51.0%, while the power levels of S_1, S_2, and S_3 are 8.4%, 10.4%, and 10.8%, respectively. When the association is weaker (OR=1.1), all tests are less powerful but the relative performance is the same, with PheProb outperforming S_1, S_2, and S_3. As expected the PheProb approach is less powerful when compared to true disease status of all individuals, but the difference in power is modest. For

example, in the setting mentioned above, the power when the true disease status is known (true-Y) is 61.6%, which is only 10.6 percentage points higher than PheProb's power of 51.0%. In most of these simulation settings, S_3 is more powerful than S_1 and S_2, but this is not always the case. For example, when the SNP OR=1.1 and the disease is uncommon (prevalence=5%), the power of S_2 is 6.4% while the power of S_3 is 5.4%.

Example 1: Real-World Clinical Study and Hyperlipidemia

Among the 1,462 patients in the lipid study, the mean age was 63.6 years and the proportion of female subjects was 80%. 64.7% of this group had at least one billing code starting with 272 (Disorders of Lipoid Metabolism); and among those with at least one billing code, the median number of billing codes was 8 and the maximum was 171. The results of the genotype-phenotype association tests are presented in Table 1. Using the thresholding method, the p-values for association between the LDL GRS and hyperlipidemia were 0.126, 0.123, and 0.142 for defining cases as having at least 1, 2, or 3 billing codes. Using the PheProb approach we observed the expected significant association between the LDL GRS and hyperlipidemia with a p-value of 0.001.

Table 1. Comparison of p-values for genotype-phenotype association tests using thresholding S_t* vs PheProb using real-world EHR data.

Phenotype	n	Genetic Marker	Phenotype-Genotype Association Test Method			
			S_1	S_2	S_3	PheProb
Hyperlipidemia (ICD-9 272.x)	1,442	LDL GRS	0.126	0.123	0.142	0.001
Rheumatoid Arthritis (ICD-9 714.x except 714.3)	14,985	HLA DRB1	<0.0001	<0.0001	<0.0001	<0.0001
		PTPN22	<0.0001	<0.0001	<0.0001	0.0002

* S_t, is the thresholding approach where subjects are defined as cases if they have ≥t PheWAS codes

Example 2: Partners Biobank and Rheumatoid Arthritis

Association between HLA DRB1 or PTPN22 and billing codes for RA

Among the 14,985 patients, 12% had at least 1 diagnosis code for RA. Among those with at least 1 billing code, the median number of billing codes was 14 and the maximum was 838. The results of the genotype-phenotype association tests are presented in Table 1. The p-values for the association between RA and the *HLA DRB1* SNP using all three thresholding methods and using the PheProb approach were less than 0.0001; the p-value for the association between the *PTPN22* and RA was less than 0.0001 using the thresholding methods and 0.0002 using the PheProb approach.

Validation

Because the p-values for all methods in this large patient population were small, we sought to better understand how well the standard thresholding methods and the PheProb clustering step were separating cases and controls by comparing accuracy metrics in a group of 546 patients for whom true RA disease status was known through chart review. To identify cases and controls from the PheProb predicted probability of having RA, we simply dichotomized it at its mean. The results are presented in Table 2. The prevalence of RA in the validation set was 8.8%. We found that PheProb better classified individuals without the disease as ``controls'' (FPR of 0.01 vs. 0.06 for S_2), and that the individuals classified by PheProb as cases were more likely to truly have RA (precision / PPV of 0.74 vs. 0.40 for S_2). This improvement in PPV came at a cost of incorrectly classifying 8% of true disease cases as controls (recall of 0.92 vs. 1.00), with no reduction in the NPV (the chance that an individual classified as control is truly disease-free; 1.00 vs. 1.00). In addition, we found that PheProb outperformed S_2 in regards to precision and recall (F1 score of 0.83 vs. 0.57, P <0.0001).

Table 2. Comparison of accuracy measures of case-control identification for RA comparing the standard thresholding method S_2* with PheProb against the true phenotype as defined by chart review.

	S2	PheProb
precision / PPV	0.40	0.74
NPV	1.00	1.00
recall	1.00	0.92
FPR	0.06	0.01

F1 score	0.57	0.83

* S_2, subjects with ≥2 PheWAS codes are defined as cases
PPV = positive predictive value; NPV = negative predictive value; FPR = false positive rate.

DISCUSSION

The PheProb approach provides a high-throughput, unsupervised method for phenotyping using existing codified data without requiring labor intensive chart review for gold standard labels. In this study, we observed that PheProb, which converts the number of diagnosis codes into a probability for a phenotype, provided more power for genetic association studies using EHR data compared to standard PheWAS thresholding approaches, while maintaining the feasibility of standard approaches by using only structured data. The thresholding approach is limited because of the varying accuracy of diagnosis codes, and PheProb addresses this challenge by using the diagnosis code counts to separate the patients into two latent underlying classes — case and controls —in a data-adaptive way while accounting for total health care utilization. This approach in effect normalizes the accuracy of the codes.

PheWASs using the thresholding method have been conducted with different threshold choices, and there does not appear to be a consensus on a best threshold [9,10,12,13,16–20,23,24]. Increasing the number of diagnosis codes used to define cases tends to increase the PPV, or the chance that those included as cases do in fact

have the disease. However, particularly for uncommon diseases, this may result in a loss of statistical power when testing for the SNP-disease association, since it reduces the number of individuals included in the test. We observed loss of power with higher thresholding in the simulations with an uncommon phenotype (prevalence=5%): with SNP OR=1.1, the power of S_2 was 6.4% while the power of S_3 was 5.4%. In this scenario, the benefit of improving the PPV of the case definition by requiring more billing codes is likely outweighed by the power lost by losing some true disease cases. Converting a quantitative variable such as the billing code count to a binary variable can also result in loss of information even if a best threshold were determined, since more codes likely reflect a more certain disease status or greater disease severity. The PheProb method avoids many of these issues by using the billing code counts directly as a quantitative variable, letting the data drive the clustering into cases and controls.

The thresholding method also does not take into account differing levels of health care utilization. We believe incorporating total health care utilization into PheProb, as quantified by the total number of billing codes, enhances performance because the amount of utilization can vary dramatically across patients. Total health care utilization can affect the informativeness of certain billing codes, particularly for diseases that may have many diagnostic tests. For example, an individual with 10 total billing codes, 5 of them for hyperlipidemia, may be more likely to be a case than an individual with 1000 total codes, 5 of which are for hyperlipidemia.

In the PheProb approach, the actual test for association is performed between the SNP and the estimated probability of having the phenotype. This eliminates the need to select a threshold to define case-control status, and carries forward into the association test some information about classification uncertainty as encoded in the estimated probability of disease. Individuals with many billing codes are assigned high probability of having the disease; those with no codes are assigned low probability; and those with a moderate number can have a probability in the middle, where "moderate" is determined in the data-driven clustering and will be slightly different for different phenotypes. Retaining this algorithmic classification uncertainty rather than thresholding has been shown to improve power in other EHR settings [33].

In the example of RA, where a gold standard was available for comparison, we found that the probability of being a case from the PheProb approach better distinguished cases and controls than the standard thresholding method. That is, identifying cases as patients with above average PheProb probability was a more accurate case definition than identifying cases as patients with more than 2 ICD-9 code counts (F1 score of 0.83 vs. 0.57, $p<0.0001$). This is likely because the clustering step seeks to separate cases and controls in a data-adaptive way, and uses both information in the disease-relevant billing codes and the total billing codes in the clustering, rather than using a fixed threshold on just the relevant billing codes.

As detailed here, many aspects of the PheProb approach were designed to improve genetic association test power: leveraging the diagnosis code counts as a quantitative variable instead of dichotomizing it; integrating healthcare utilization

in the structure of the mixture model; and using the continuous model-predicted probability of being a disease case as an outcome instead of thresholding it to identify case-control status. Unfortunately, it is difficult to disentangle the impacts on power of each of these aspects. For example, a patient's healthcare utilization C enters the method as a parameter in the binomial mixture model; avoiding reliance on C would necessitate using a different parametric mixture model, but performance differences due to the change in model and the reliance on C cannot be rigorously disentangled. Similarly, we propose using the model-predicted probability rather than a dichotomous outcome in the genetic association test; this is because we feel that the probability better carries forward phenotyping uncertainty into the test, but it is also not obvious how best to dichotomize that predicted probability into a case-control outcome in the absence of any gold-standard outcome data, and whether patients with mid-level predicted probability should be excluded in such a dichotomization.

With no additional covariates (such as age and gender), our genotype-phenotype test is essentially model-free — it is fundamentally testing whether the genotype and phenotype are statistically independent. Thus, it is a valid test across a wide range of true disease-genotype models. With additional covariates, our model is still valid across a wide range of true disease-genotype models so long as the genotype is independent of those covariates. Relaxing this assumption to produce an even more robust test is a direction of future research.

The PheProb method is designed to increase power for studies which rely on (or are limited to using) ICD codes for phenotyping, with direct applications to

PheWAS. It was designed for studies screening across hundreds to thousands of phenotypes, where creating individual highly accurate algorithms for each phenotype is not feasible. While it may serve as a starting point for investigators interested in detailed studies on a specific phenotype, it was not designed for this purpose.

Institutions have different EHR systems and different approaches to using those systems. As a result, phenotyping methods typically have varying performance across institutions. This variability is especially notable when phenotyping is based only on thresholded diagnosis codes. For example, using at least one code for RA to identify cases was shown to have a PPV of 22%, 26%, and 49% across three institutions, while using at least three codes was shown to have a PPV of 55%, 42%, and 73% across the same three institutions [21]. This highlights another difficulty of existing PheWAS methods based on thresholding — even for a single disease, the best threshold for defining a case may vary across institutions. A method like PheProb that has the potential to adapt to the underlying distribution of diagnosis codes at an institution may be effective in the face of this heterogeneity, and further evaluating PheProb's performance and robustness across healthcare systems is a direction of future research.

CONCLUSION

Compared to the standard PheWAS approach which defines phenotypes using a certain number of diagnosis codes, the PheProb approach provides more power to study genotype-phenotype associations by retaining information on the count of

billing codes defining the phenotype normalized by health care utilization information, and providing a probability of a phenotype rather than a binary case-control status for association testing. In sum, PheProb is a scalable method for rapid unsupervised phenotyping with direct applications for PheWAS and large scale EHR-biorepository genetic association studies.

FIGURE LEGENDS

Figure 1. Workflow of the PheProb method. True disease status of patients is unknown; instead, the number of billing codes for each disease is observed. The PheProb method clusters individuals based on billing codes, and tests for an association between a genetic marker such as a SNP and the clustering-based probability of being a case.

Figure 2. Comparison of power to detect an association between a SNP and a phenotype (a disease) when only S, the count of billing codes, is observed. The standard approaches of thresholding S and identifying disease cases as individuals with more than t billing codes are indicated by S_t, for t=1,2,3. Our proposed method is indicated by PheProb. A benchmark method is also shown: true-Y, which uses the true disease status as the outcome. Simulation settings vary the prevalence and the strength of the association of interest between disease status and the SNP: OR=1.1 for a weak relationship and OR=1.35 for a moderate relationship.

Funding Statement: This work was supported by United States National Institutes of Health grants U54-HG007963, U54-LM008748, R01-HL089778, R01-HL127118, F31GM119263-01A1, K23-DK097142, P30-AR072577, the Harold and Duval Bowen Fund, and internal funds from Tsinghua University and Partners HealthCare.

Competing Interests Statement: The authors have no competing interests to declare.

Contributorship Statement: All authors made substantial contributions to: conception and design; acquisition, analysis and interpretation of data; drafting the article or revising it critically for important intellectual content; and final approval of the version to be published.

References

1 Kohane IS. Using electronic health records to drive discovery in disease genomics. *Nat Rev Genet* 2011;**12**:417–428.

2 Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *J Am Med Inform Assoc* 2013;:e206–e211.

3 Ritchie MD, Denny JC, Crawford DC, *et al.* Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet* 2010;**86**:560–572.

4 Kurreeman F, Liao K, Chibnik L, *et al.* Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. *Am J Hum Genet* 2011;**88**:57–69.

5 Kho AN, Hayes MG, Rasmussen-Torvik L, *et al.* Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J Am Med Inform Assoc* 2011;**19**:212–218.

6 Xu H, Jiang M, Oetjens M, *et al.* Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. *J Am Med Inform Assoc* 2011;**18**:387–391.

7 Jones R, Pembrey M, Golding J, *et al.* The search for genenotype/phenotype associations and the phenome scan. *Paediatr Perinat Epidemiol* 2005;**19**:264–275.

8 Bilder RM, Sabb FW, Cannon TD, *et al.* Phenomics: the systematic study of phenotypes on a genome-wide scale. *Neuroscience* 2009;**164**:30–42.

9 Denny JC, Ritchie MD, Basford MA, *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* 2010;**26**:1205–1210.

10 Hebbring SJ, Schrodi SJ, Ye Z, *et al.* A PheWAS approach in studying HLA-DRB1*1501. *Genes Immun* 2013;**14**:187–91. doi:10.1038/gene.2013.2

11 Namjou B, Marsolo K, Caroll RJ, *et al.* Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts, genetically links PLCL1 to speech language development and IL5-IL13 to Eosinophilic Esophagitis. *Front Genet*

2014;**5**.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4235428/ (accessed 7 Aug 2017).

12 Shameer K, Denny JC, Ding K, *et al.* A genome-and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. *Hum Genet* 2014;**133**:95–109.

13 Cronin RM, Field JR, Bradford Y, *et al.* Phenome-wide association studies demonstrating pleiotropy of genetic variants within FTO with and without adjustment for body mass index. *Front Genet* 2014;**5**.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4134007/ (accessed 7 Aug 2017).

14 Mitchell SL, Hall JB, Goodloe RJ, *et al.* Investigating the relationship between mitochondrial genetic variation and cardiovascular-related traits to develop a framework for mitochondrial phenome-wide association studies. *BioData Min* 2014;**7**:6. doi:10.1186/1756-0381-7-6

15 Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinforma Oxf Engl* 2014;**30**:2375–6. doi:10.1093/bioinformatics/btu197

16 Diogo D, Bastarache L, Liao KP, *et al.* TYK2 protein-coding variants protect against rheumatoid arthritis and autoimmunity, with no evidence of major pleiotropic effects on non-autoimmune complex traits. *PLoS One* 2015;**10**:e0122271.

17 Verma A, Basile AO, Bradford Y, *et al.* Phenome-Wide Association Study to Explore Relationships between Immune System Related Genetic Loci and Complex Traits and Diseases. *PLOS ONE* 2016;**11**:e0160573. doi:10.1371/journal.pone.0160573

18 Evidence for extensive pleiotropy among pharmacogenes. *Pharmacogenomics* 2016;**17**:853–66. doi:10.2217/pgs-2015-0007

19 Liu J, Ye Z, Mayer JG, *et al.* Phenome-wide association study maps new diseases to the human major histocompatibility complex region. *J Med Genet* 2016;:jmedgenet-2016-103867. doi:10.1136/jmedgenet-2016-103867

20 Karnes JH, Bastarache L, Shaffer CM, *et al.* Phenome-wide scanning identifies multiple diseases and disease severity phenotypes associated with HLA variants. *Sci Transl Med* 2017;**9**. doi:10.1126/scitranslmed.aai8708

21 Carroll RJ, Thompson WK, Eyler AE, *et al.* Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *J Am Med Inform Assoc* 2012;**19**:e162–9. doi:10.1136/amiajnl-2011-000583

22 Leader JB, Pendergrass SA, Verma A, *et al.* Contrasting Association Results between Existing PheWAS Phenotype Definition Methods and Five Validated Electronic Phenotypes. In: *AMIA Annual Symposium Proceedings*. American Medical Informatics Association 2015. 824.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4765620/ (accessed 7 Aug 2017).

23 Denny JC, Bastarache L, Ritchie MD, *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* 2013;**31**:1102.

24 Ritchie MD, Denny JC, Zuvich RL, *et al.* Genome-and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. *Circulation* 2013;**127**:1377–1385.

25 Leisch F. Flexmix: A general framework for finite mixture models and latent glass regression in R. Published Online First: 2004.http://ro.uow.edu.au/buspapers/487/

26 Grün B, Leisch F. Fitting finite mixtures of generalized linear regressions in R. *Comput Stat Data Anal* 2007;**51**:5247–5252.

27 Grun B, Leisch F. FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. Published Online First: 2008.http://ro.uow.edu.au/commpapers/1140/

28 Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B Methodol* 1977;:1–38.

29 Carey VJ, Lumley T, Ripley BD. gee: Generalized estimation equation solver. *UR HttpCRAN R-Proj Orgpackage Gee R Package Version* 2012;:4–13.

30 Team RC. *A language and environment for statistical computing. R Foundation for statistical computing, 2015; Vienna, Austria*. 2016.

31 Liao KP, Diogo D, Cui J, *et al.* Association between low density lipoprotein and rheumatoid arthritis genetic factors with low density lipoprotein levels in rheumatoid arthritis and non-rheumatoid arthritis controls. *Ann Rheum Dis* 2014;**73**:1170–1175.

32 Gainer VS, Cagan A, Castro VM, *et al.* The Biobank Portal for Partners Personalized Medicine: A Query Tool for Working with Consented Biobank Samples, Genotypes, and Phenotypes Using i2b2. *J Pers Med* 2016;**6**:11. doi:10.3390/jpm6010011

33 Sinnott JA, Dai W, Liao KP, *et al.* Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records. *Hum Genet* 2014;**133**:1369.